On the Feasibility and Robustness of Pointwise Evaluation of QPP

Suchana Datta

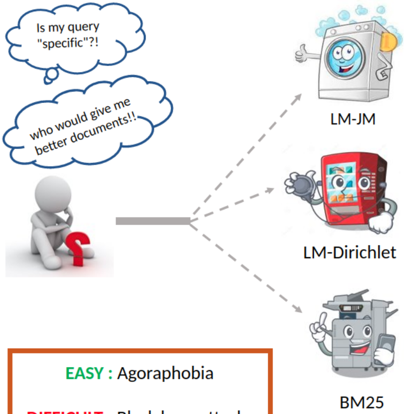
University College Dublin

Debasis Ganguly (University of Glasgow)

Derek Greene (University College Dublin)

Mandar Mitra (Indian Statistical Institute)

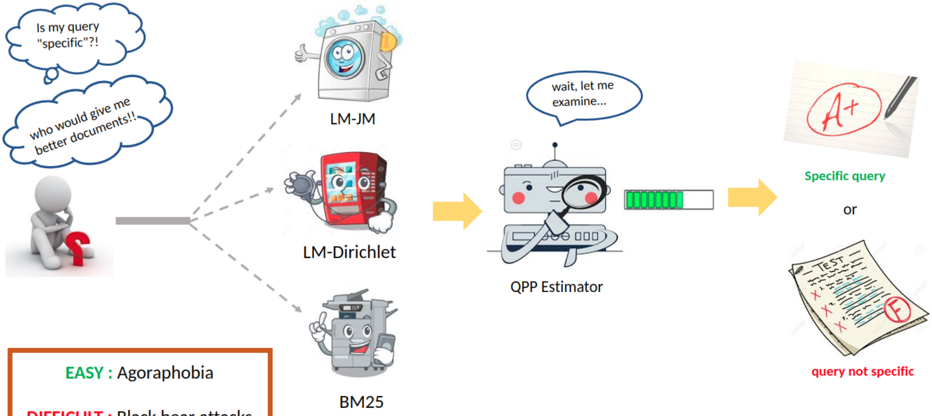
What is Query Performance Prediction (QPP)?



EASY : Agoraphobia

DIFFICULT : Black bear attacks

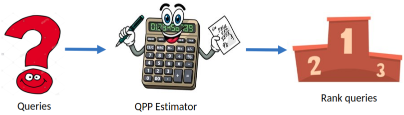
What is Query Performance Prediction (QPP)?



EASY : Agoraphobia

DIFFICULT : Black bear attacks

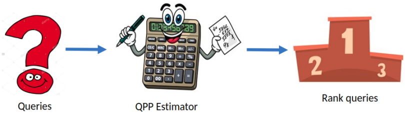
Types of QPP estimators

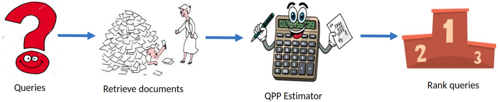


Pre-retrieval

- Predicts the performance of each query based on the content and the context of the query.
- Predictors are often derived from linguistic or statistical information.
- AvgIDF, MaxIDF.

Types of QPP estimators





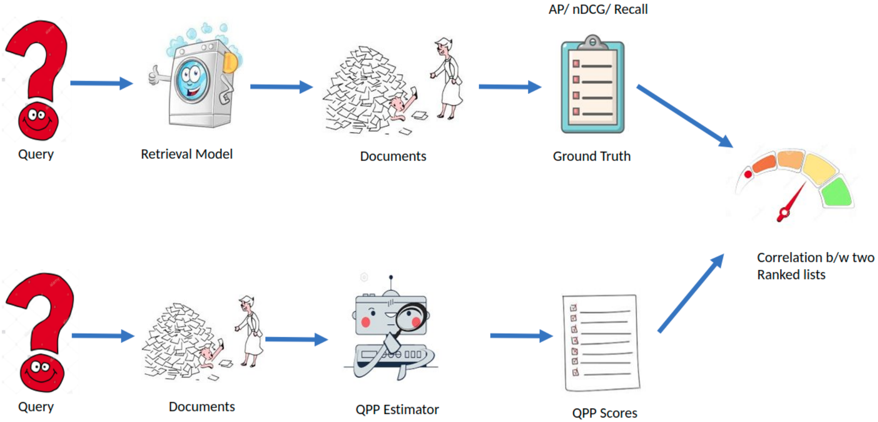
Pre-retrieval

- Predicts the performance of each query based on the content and the context of the query.
- Predictors are often derived from linguistic or statistical information.
- AvgIDF, MaxIDF.

Post-retrieval

- Estimates the query performance by analyzing the result list returned by the retrieval engine.
- Clarity-based approaches - Clarity.
- Score-based approaches - WIG, NQC.
- Robustness-based approaches - UEF.

How do we evaluate QPP estimators?



Do we really want 'listwise' evaluation?

- Performance of one query is measured relative to the others.

Do we really want 'listwise' evaluation?

- Performance of one query is measured relative to the others.
- A downstream performance estimate of an individual query also needs to be evaluated independently.

Do we really want 'listwise' evaluation?

- Performance of one query is measured relative to the others.
- A downstream performance estimate of an individual query also needs to be evaluated independently.
- A pointwise approach measures the effectiveness on individual queries.

Do we really want 'listwise' evaluation?

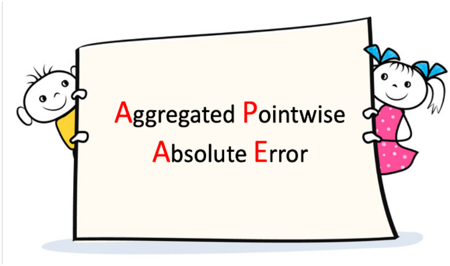
- Performance of one query is measured relative to the others.
- A downstream performance estimate of an individual query also needs to be evaluated independently.
- A pointwise approach measures the effectiveness on individual queries.
- Allows us to carry out a per-query analysis of a method.

Do we really want ‘listwise’ evaluation?

- Performance of one query is measured relative to the others.
- A downstream performance estimate of an individual query also needs to be evaluated independently.
- A pointwise approach measures the effectiveness on individual queries.
- Allows us to carry out a per-query analysis of a method.
- Listwise methods can be overly sensitive to the configuration setup used for evaluation[a].

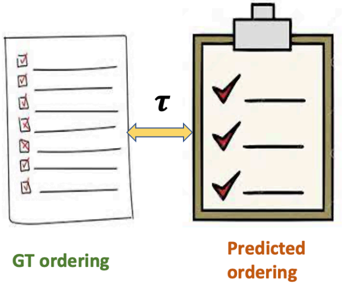
[a] D. Ganguly, S. Datta, M. Mitra, D. Greene, An analysis of variations in the effectiveness of query performance prediction, in: Proc. of ECIR' 22, 2022, pp. 215–229.

What do we propose? - APAE

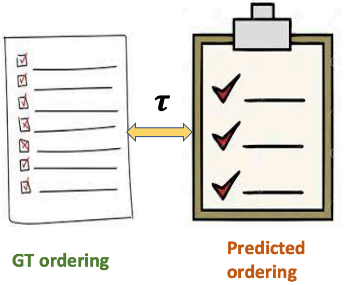


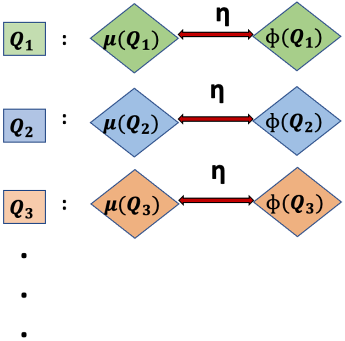
- A new QPP evaluation framework.
- Shown to be **consistent** with the existing listwise approaches.
- **More robust** to changes in QPP experimental setup.

Individual ground-truth



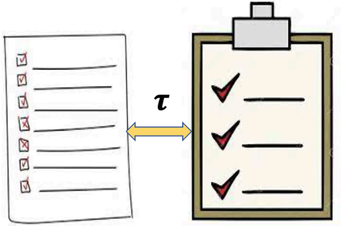
Individual ground-truth

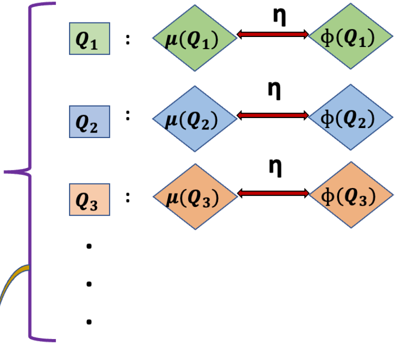




$$\eta(\mu(Q), \varphi(Q)) = 1 - |\mu(Q) - \varphi(Q) / \aleph|$$

Individual ground-truth

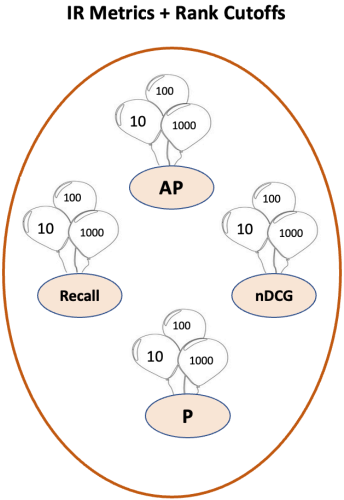




$$\eta(\mu(Q), \varphi(Q)) = 1 - |\mu(Q) - \varphi(Q) / \aleph|$$

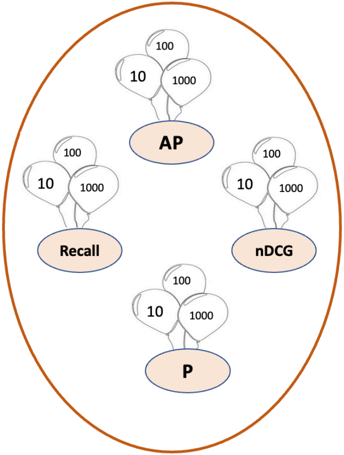
$$\boxed{\eta(\mathcal{Q}, \mu, \varphi) = \frac{1}{|\mathcal{Q}|} \Sigma_{Q \in \mathcal{Q}}\, \eta(\mu(Q), \varphi(Q))}$$

Metric-agnostic pointwise QPP evaluation



Metric-agnostic pointwise QPP evaluation

IR Metrics + Rank Cutoffs



- $\eta(Q, \mathcal{M}, \varphi) = \Sigma_{\mu \in \mathcal{M}} (1 - |\mu(Q) - \varphi(Q) / \aleph|)$

$$\boxed{\Sigma \in \{\text{avg, min, max}\}}$$

- Average over these values for a set of queries.

- $$\boxed{\eta(\mathcal{Q}, \mu, \varphi) = \frac{1}{|\mathcal{Q}|} \Sigma_{Q \in \mathcal{Q}} \eta(\mu(Q), \varphi(Q))}$$

We investigate -

RQ1: Does APAE **agree** with the standard listwise correlation metrics?

We investigate -

RQ1: Does APAE **agree** with the standard listwise correlation metrics?

RQ2: How **robust** is APAE with respect to changes in the QPP experiment context?

We investigate -

RQ1: Does APAE **agree** with the standard listwise correlation metrics?

RQ2: How **robust** is APAE with respect to changes in the QPP experiment context?

Dataset : TREC Robust - 249 queries

Observations in relation to RQ1

RQ1: Does APAE **agree** with the standard listwise correlation metrics?

	$\eta_{\text{avg}}(\mathcal{M})$				$\eta_{\text{min}}(\mathcal{M})$				$\eta_{\text{max}}(\mathcal{M})$			
	r	ρ	τ	sARE	r	ρ	τ	sARE	r	ρ	τ	sARE
BM25	0.810	0.810	0.905	0.887	0.778	0.778	0.794	0.813	0.802	0.810	0.794	0.794
LMDir	**0.905**	**0.810**	**0.905**	**0.887**	0.778	0.794	0.794	0.810	0.769	0.782	0.794	0.796
LMJM	0.810	0.810	0.810	0.846	0.794	0.794	0.782	0.786	0.794	0.769	0.810	0.846

Observations in relation to RQ2

RQ2: How **robust** is APAE with respect to changes in the QPP experiment context?

Model	Metric	AP@100	R@10	R@100	nDCG@10	nDCG@100
LMJM		0.497	0.813	0.429	0.783	0.429
BM25	AP@10	0.897	0.722	0.722	0.793	0.793
LMDir		0.897	0.786	0.786	0.823	0.905
LMJM			**0.328**	0.811	0.363	0.783
BM25	AP@100		0.783	0.784	0.714	0.642
LMDir			0.823	0.901	0.834	0.789
LMJM				0.624	0.893	0.503
BM25	R@10			0.803	0.982	0.894
LMDir				0.903	0.864	0.864
LMJM					0.852	0.804
BM25	R@100				0.786	0.890
LMDir					0.738	0.738
LMJM						0.537
BM25	nDCG@10					0.904
LMDir						0.868

Model	Metric	AP@100	R@10	R@100	nDCG@10	nDCG@100
LMJM		0.904	1.000	0.715	1.000	0.792
BM25	AP@10	1.000	1.000	1.000	1.000	1.000
LMDir		1.000	1.000	1.000	1.000	1.000
LMJM			0.905	0.811	0.669	1.000
BM25	AP@100		1.000	1.000	1.000	1.000
LMDir			1.000	1.000	1.000	1.000
LMJM				0.603	0.905	**0.542**
BM25	R@10			1.000	1.000	1.000
LMDir				1.000	1.000	1.000
LMJM					0.654	1.000
BM25	R@100				1.000	1.000
LMDir					1.000	1.000
LMJM						0.649
BM25	nDCG@10					1.000
LMDir						1.000

Observations in relation to RQ2

RQ2: How **robust** is APAE with respect to changes in the QPP experiment context?

Metric	Model	LMJM (0.6)	BM25 (0.7, 0.3)	BM25 (0.3, 0.7)	LMDir (500)	LMDir (1000)
AP@100		0.826	0.904	0.819	0.714	0.895
nDCG@100	LMJM	0.780	0.694	0.695	0.759	0.759
R@100	(0.3)	0.824	0.769	0.782	0.904	0.904
AP@100			0.703	0.712	0.904	0.823
nDCG@100	LMJM		0.781	0.827	0.811	0.811
R@100	(0.6)		0.813	0.725	0.731	**0.675**
AP@100				0.903	0.785	0.785
nDCG@100	BM25			0.897	0.786	0.786
R@100	(0.7, 0.3)			0.812	0.752	0.779
AP@100					0.887	0.882
nDCG@100	BM25				0.901	0.895
R@100	(0.3, 0.7)				0.889	0.901
AP@100						0.901
nDCG@100	LMDir					0.893
R@100	(500)					0.903

Metric	Model	LMJM (0.6)	BM25 (0.7, 0.3)	BM25 (0.3, 0.7)	LMDir (500)	LMDir (1000)
AP@100		1.000	1.000	1.000	1.000	1.000
nDCG@100	LMJM	1.000	0.864	1.000	0.843	0.864
R@100	(0.3)	1.000	0.864	1.000	1.000	1.000
AP@100			1.000	1.000	1.000	1.000
nDCG@100	LMJM		0.914	1.000	**0.813**	0.914
R@100	(0.6)		1.000	1.000	1.000	1.000
AP@100				1.000	1.000	1.000
nDCG@100	BM25			1.000	1.000	1.000
R@100	(0.7, 0.3)			0.812	0.905	1.000
AP@100					1.000	1.000
nDCG@100	BM25				1.000	1.000
R@100	(0.3, 0.7)				1.000	1.000
AP@100						1.000
nDCG@100	LMDir					1.000
R@100	(500)					1.000

Concluding Remarks

- We propose a pointwise evaluation method that computes the relative difference between a normalized QPP score and a true IR evaluation measure.
- The proposed metric exhibits a high correlation with standard listwise approaches.
- More robust to changes in QPP experimental setup than listwise evaluation measures.
- It is possible to evaluate the effectiveness of different QPP methods on downstream tasks on a per-query basis.

Thank you for your attention!!